|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subject Title: INSIGHTS OF BIG DATA** | | | | | |
| **Course Code: PE4101** | **Year and Semester:** IV Year I semester | L | T | P | C |
| **Prerequisites:** Basic Knowledge on   * Java Programming * RDBMS * Data MIning | | 4 | 0 | 0 | 3 |

**Course Objectives:**

1. To understand the complexity and volume of Big Data and their challenges.
2. To analyze the various methods of data collection.
3. To comprehend the necessity for pre-processing Big Data and their issues.
4. To understand predictive analytics and descriptive analytics.
5. To understand and implement Big Data Analytics with data convergence and Business Maturity Model.

**Course Outcomes:** At the end of the course student will be able to:

|  |  |
| --- | --- |
| **COS’** | **STATEMENT** |
| **CO1** | **Identify** the various sources of Big Data. |
| **CO2** | **List** the components of Hadoop and Hadoop Eco-System and Analyze file systems such as GFS and HDFS. |
| **CO3** | **Apply** map reduce concepts for desired applications. |
| **CO4** | **Demonstrate** the Pig architecture and evaluation of pig scripts. |
| **CO5** | **Implement** Big Data Activities using Hive |

**UNIT - I 9 Hours**

**INTRODUCTION TO BIG DATA:** Data, Characteristics of data and types of digital data, Sources of data, Working with unstructured data, Evolution and definition of big data, Characteristics and need of big data, Challenges of big data.

**BIG DATA ANALYTICS:** Overview of business intelligence, Data science and analytics, Meaning and characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment.

**UNIT - II 9 Hours**

**INTRODUCTION TO HADOOP :** Introducing hadoop, Need of hadoop, Limitations of RDBMS, RDBMS versus hadoop, Distributed computing challenges, History of hadoop , Hadoop overview, Use case of hadoop, Hadoop distributors, HDFS (Hadoop distributed file system), Processing data with hadoop, Managing resources and applications with hadoop YARN (yet another resource negotiator), Interacting with hadoop ecosystem.

**UNIT - III 9 Hours**

**INTRODUCTION TO MAPREDUCE PROGRAMMING**: Introduction-mapper, reducer, combiner, partitioner, searching, sorting, compression, real time applications using mapreduce, combiner, partitioner, matrix multiplication using mapreduce and page rank algorithm using mapreduce.

**UNIT - IV 9 Hours**

**INTRODUCTION TO PIG**: The anatomy of pig, Pig on hadoop, Pig philosophy, Usecase for pig, ETL processing, Pig latin overview, Data types in pig, Running pig, Execution modes of pig, HDFS commands, Relational operators, Piggy bank, Word count example using pig, Pig at Yahoo.

**UNIT - V 9 Hours**

**INTRODUCTION TO HIVE:** Introduction to hive, Hive architecture, Hive data types, Hive file format, Hive query language (HQL).

HIVE: Partitions and bucketing, RCFile Implementation, working with XML files, User-defined Function (UDF) in Hive, Pig versus Hive.

**Text Books:**

1. Seema Acharya, Subhashini Chellappan, “Big Data and Analytics”, 1st edition, Wiley, Publishers, 2015.

**Reference Books:**

1. Boris lublinsky, Kevin t. Smith, AlexeyYakubovich, “Professional Hadoop Solutions”,

1st edition, Wiley, 2015.

2.Chris Eaton, Dirkderoosetal, “Understanding Big data “, 1st edition, McGraw Hill, 2012.

3.Tom White, “HADOOP: The definitive Guide”, 1st edition, O Reilly 2012.

4.Vignesh Prajapati, “Big Data Analytics with R and Hadoop”,1st edition, Packet Publishing,

2013.

**Software Links:**

1.Hadoop: http://hadoop.apache.org/

2.Hive:https://cwiki.apache.org/confluence/display/Hive/HomePigLatin: http://pig.apache.org/docs/r0.7.0/tutorial.html

**Micro Syllabus of Insights of Big Data**

**IV B. Tech I Semester**

|  |  |  |
| --- | --- | --- |
| **UNIT I : INTRODUCTION TO BIG DATA:** Data, Characteristics of data and types of digital data, Sources of data, Working with unstructured data, Evolution and definition of big data, Characteristics and need of big data, Challenges of big data.  **BIG DATA ANALYTICS:** Overview of business intelligence, Data science and analytics, Meaning and characteristics of big data analytics, Need of big data analytics, Classification of analytics, Challenges to big data analytics, Importance of big data analytics, Basic terminologies in big data environment. | | |
| **Unit** | **Module** | **Micro Content** |
| UNIT I | Introduction to Big Data | Data |
| Characteristics of data |
| Types of digital data |
| Sources of data |
| Working with unstructured data |
| Evolution and Definition of big data |
| Characteristics and Need of big data |
| Challenges of big data |
| Evolution and Definition of big data |
| Characteristics and Need of big data |
| Big Data Analytics | Overview of business intelligence |
| Data science and Analytics |
| Meaning and Characteristics of big data analytics |
|  | Need of big data analytics |
|  | Classification of analytics |
|  | Challenges to big data analytics |
|  | Importance of big data analytics |
|  | Basic terminologies in big data environment |
| **UNIT – II: INTRODUCTION TO HADOOP :** Introducing hadoop, Need of hadoop, Limitations of RDBMS, RDBMS versus hadoop, Distributed computing challenges, History of hadoop , Hadoop overview, Use case of hadoop, Hadoop distributors, HDFS (Hadoop distributed file system), Processing data with hadoop, Managing resources and applications with hadoop YARN (yet another resource negotiator), Interacting with hadoop ecosystem. | | |
| **Unit** | **Module** | **Micro Content** |
| UNIT II | Basics of Hadoop | Introducing Hadoop |
| Need of Hadoop |
| Limitations of RDBMS |
| RDBMS versus Hadoop |
| Distributed Computing Challenges |
| History of Hadoop |
| Hadoop Overview | Hadoop Overview |
|  | Hadoop distributors |
|  | Use case of hadoop |
|  | Processing data with hadoop |
|  | Hadoop distributed file system |
|  | Hadoop Eco System |
|  | YARN Yet Another Resource Negotiator |
| **UNIT – III : INTRODUCTION TO MAPREDUCE PROGRAMMING**: Introduction-mapper, reducer, combiner, partitioner, searching, sorting, compression, real time applications using mapreduce, combiner, partitioner, matrix multiplication using mapreduce and page rank algorithm using mapreduce. | | |
| **Unit** | **Module** | **Micro Content** |
| UNIT III | MapReduce Programming | Introduction |
| Maper |
| Reducer |
| Combiner |
| Partitioner |
| Searching |
| Sorting |
| Compression |
| real time applications using mapreduce |
| matrix multiplication using mapreduce |
| page rank algorithm using mapreduce |
| **UNIT - IV : INTRODUCTION TO PIG**: The anatomy of pig, Pig on hadoop, Pig philosophy, Usecase for pig, ETL processing, Pig latin overview, Data types in pig, Running pig, Execution modes of pig, HDFS commands, Relational operators, Piggy bank, Word count example using pig, Pig at Yahoo. | | |
| **Unit**  UNIT IV | **Module** | **Micro Content** |
| Introduction and Overview of Pig | PIG |
| Motivation of using pig |
| The anatomy of pig |
| Key Features of Pig |
| Pig on Hadoop |
| Pig Philosophy |
| Usecase for pig -ETL Processing |
| Pig Latin Overview |
| Pig Data Types |
| Running Pig |
| Execution Modes of Pig |
| HDFS Commands |
| Pig Operations |
| Piggy Bank |
| Word count example using pig |
| Pig at Yahoo |
| **UNIT V : INTRODUCTION TO HIVE:** Introduction to hive, Hive architecture, Hive data types, Hive file format, Hive query language (HQL)  HIVE: Partitions and bucketing, RCFile Implementation, working with XML files, User-defined Function (UDF) in Hive, Pig versus Hive. | | |
| **Unit** | **Module** | **Micro Content** |
| UNIT V | Hive | Introduction to Hive |
| Hive architecture |
| Hive Data Types |
| Hive File Foramt |
| Hive Query Language(HQL) |
| Hive Partitions and bucketing |
| RCFile Implementatiom |
| Working with XML files |
| User defined Function(UDF) in Hive |
| Pig versus Hive |

**CO – PO MAPPING**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **POs** | | | | | | | | | | | | | |
| **CO/ PO** | **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **P011** | **PO12** | **PSO1** | **PSO2** |
| **CO1** | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **CO2** | 3 | 2 |  |  |  |  |  |  |  |  |  |  |  |  |
| **CO3** | 3 | 3 | 3 |  |  |  |  |  |  |  |  |  |  |  |
| **CO4** | 3 | 3 | 3 | 3 |  |  |  |  |  |  |  |  |  |  |
| **CO5** | 3 | 3 | 3 |  |  |  |  |  |  |  |  |  |  |  |

**Abridged CO-PO MAPPING TABLE**

|  |  |  |
| --- | --- | --- |
| **CO’S** | **STATEMENT** | **PO’s** |
| **CO1** | Identify the various sources of Big Data. | 1 |
| **CO2** | List the components of Hadoop and Hadoop Eco-System and Analyze file systems such as GFS and HDFS. | 1,2 |
| **CO3** | Apply map reduce concepts for desired applications. | 1,2,3 |
| **CO4** | Demonstrate the Pig architecture and evaluation of pig scripts. | 1,2,3,4 |
| **CO5** | Implement Big Data Activities using Hive | 1,2,3 |

**Justification:**

|  |  |  |  |
| --- | --- | --- | --- |
| **CO No** | **PO/**  **PSO** | **CL** | **Justification** |
| 1 | 1 | 3 | Strongly having the Knowledge of the fundamental concepts of Big Data management and analytics helps in solving complex engineering problems. |
| 2 | 1 | 3 | Slightly having the knowledge of Big Data management and analytics concepts knowledge can be used to design and conduct experiments to provide valid conclusions |
| 2 | 2 | Moderately the student will know Principles of mathematics and engineering sciences are used in various aspects of data analytic approaches. |
| 3 | 1 | 3 | Strongly having Knowledge of Map Reduce paradigm and the Apache Hadoop system involves solving complex engineering problems |
| 2 | 3 | Strongly having Knowledge of Map Reduce paradigm and the Hadoop system knowledge can be used to conduct experiments in real life problems to provide valid conclusions. |
| 3 | 3 | Strongly mapped design solutions for complex engineering problems |
| 4 | 1 | 3 | Strongly mapped student will deep look into how Apache Pig works and data analysis with an example data set. |
| 2 | 3 | Analyze larger sets of data representing them as data flows |
| 3 | 3 | Deal with a large amount of data for analysis without writing any MapReduce jobs or scripts. |
| 4 | 3 | Propose a solution to process data from multiple sources, such as HBase, Hive, etc. |
| 5 | 1 | 3 | Strongly mapped to functionality of reading, writing, and managing large datasets residing in distributed storage. |
| 2 | 3 | Strongly mapped student can analyze structured data |
| 3 | 3 | Strongly student can design a solution for querying and managing only structured data stored in tables. |

COURSE COORDINATER MODULE COORDINATER HOD